Multi-modal Deep Learning for Fuji Apple Detection Using RGB-D Cameras and their Radiometric Capabilities

Jordi Gené-Mola[a], Verónica Vilaplana[b], Joan R. Rosell-Polo[a], Josep-Ramon Morros[b], Javier Ruiz-Hidalgo[b], Eduard Gregorio[a,*]

[a]*Research Group in AgroICT & Precision Agriculture, Department of Agricultural and Forest Engineering, Universitat de Lleida (UdL) – Agrotecnio Center, Lleida, Catalonia, Spain.*

[b]*Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, Barcelona, Catalonia, Spain.*

Abstract

Fruit detection and localization will be essential for future agronomic management of fruit crops, with applications in yield prediction, yield mapping and automated harvesting. RGB-D cameras are promising sensors for fruit detection given that they provide geometrical information with color data. Some of these sensors work on the principle of time-of-flight (ToF) and, besides color and depth, provide the backscatter signal intensity. However, this radiometric capability has not been exploited for fruit detection applications. This work presents the KFuji RGB-DS database, composed of 967 multi-modal images containing a total of 12,839 Fuji apples. Compilation of the database allowed a study of the usefulness of fusing RGB-D and radiometric information obtained with Kinect v2 for fruit detection. To do so, the signal intensity was range corrected to overcome signal attenuation, obtaining an image that was proportional to the reflectance of the scene. A registration between RGB, depth and intensity images was then carried out. The Faster R-CNN model was adapted for use with five-channel input images: color (RGB), depth (D) and range-corrected intensity signal (S). Results show an improvement of 4.46% in F1-score when adding depth and range-corrected intensity channels, obtaining an F1-score of 0.898 and an AP of 94.8% when all channels are used. From our experimental results, it can be concluded that the radiometric capabilities of ToF sensors give valuable information for fruit detection.

Keywords: RGB-D; Multi-modal faster R-CNN; Convolutional Neural Networks; Fruit detection; Agricultural robotics; Fruit reflectance

1. Introduction

To meet the food needs of a world's growing population, horticulture must find new ways to increase the production of fruits and vegetables (Siegel et al., 2014). This is a major challenge for agricultural communities, especially in a context of rising farming costs and a shortage of skilled labor. Efficient and sustainable agronomic management is required to reduce economic and environmental costs while increasing orchard productivity.

* *Corresponding author.*

E-mail address: egregorio@eagrof.udl.cat

Improvements in technological fields like robotics and computer science have provided farmers with tools to increase production in an efficient and sustainable way (Underwood et al., 2016). The use of new technologies in precision agriculture has been applied in the optimization of agricultural processes such as water irrigation, agrochemical application, fertilization, pruning and thinning (Auat Cheein and Carelli, 2013; Bargoti and Underwood, 2017b). Farmers can obtain valuable information for optimization of these processes from the detection and quantification of fruit distribution within the canopy.

Advances in sensing and computer vision have facilitated the development of remote fruit detection systems, with applications in yield prediction, yield mapping and automated harvesting. Yield prediction allows farmers to plan the harvest campaign, fruit storage and sales (Bargoti and Underwood, 2017b; Nuske et al., 2014). On many occasions, yield estimation is carried out by manual counting of a few samples, without addressing spatial variability within the orchard. Although simple random sampling (SRS) is a widely used technique for yield estimation, it is necessary to sample a relatively large number of trees for a precise estimation. Though this may sometimes be unfeasible with manual counting, it could be possible by using currently available computer vision technologies. As for yield mapping, production maps provide useful information for fruit growers. Fruit orchards usually show spatial variability due to soil variations, fertility, water irrigation, among others (Uribeetxebarria et al., 2018). An analysis of yield maps helps farmers to find the reasons for such variability and to determine which areas of lower productivity require special attention. Finally, fruit detection and 3D localization are the first steps in the development of automated harvesting. Hand harvesting is a hard and human-resource intensive labor, which has to find an alternative since the decreasing availability of skilled labor force (Gongal et al., 2015). Despite the latest advances in imaging techniques and computer vision, detecting and localizing fruits within the canopy is still a pending issue that has to face problems derived from the heterogeneity of the environment, such as occlusions with other vegetative organs and variable lighting conditions. Most of the emerging sensors, such depth cameras (RGB-D sensors), have not yet been exploited for fruit detection and localization. The major reason is the lack of substantial datasets (Hameed et al., 2018).

This paper introduces the KFuji RGB-DS database, which contains multi-modal images of Fuji apples in real orchards, and presents a novel study of the usefulness of RGB-D sensors and their radiometric capabilities for fruit detection. The Faster Region-based Convolutional Neural Network (Faster R-CNN) was adapted and implemented for apple detection using multi-modal images obtained with Microsoft's Kinect v2 (Microsoft, Redmond, WA, USA). The multi-modal images were obtained after pre-processing and registering three different modalities: color (RGB), depth (D) and range-corrected IR intensity -proportional to reflectance- (S).

The main contributions of this paper are: (1) provision of the first apple dataset with multi-modal images from RGB-D sensors with color, depth and range-corrected IR intensity data, and the corresponding annotations with the ground truth apple locations; (2) an analysis of the radiometric capabilities of Kinect v2 for fruit detection; (3) an implementation of a high-performance fruit detection system using an adaptation of Faster R-CNN for five-channel input images; (4) a study of the optimal anchor scales and aspect ratios used in the region proposal network (RPN). After this Introduction section, the rest of the paper is structured as follows: section 2 presents related work retrieved from the state of the art; section 3 describes the proposed dataset, explaining the signal range-correction theoretical basis, the experimental set up for data acquisition, the pre-processing needed to build the 5-channel multi-modal images, and the network implemented for fruit detection; section 4 shows the results and discusses qualitatively and quantitatively the performance of the fruit detector when using each of the modalities provided by the sensor; finally, the conclusions are presented in section 5.

.

2. Related work

Over the years, different sensors and systems have been used for fruit detection and localization (Gongal et al., 2015). The most commonly sensors used are color (or RGB) cameras (Bargoti and Underwood, 2017a; Linker, 2017; Maldonado and Barbosa, 2016; Zhao et al., 2016). However, the drawbacks to these sensors include the fact that they only provide 2D information and their measurements are affected by lighting conditions. Advances in photonics and the exploration of non-visible wavelengths have allowed the introduction of other systems, including thermal, multispectral and hyperspectral cameras. Thermal cameras have been used in fruit detection, differentiating fruits from background by the different thermal inertia of the fruits. Fruits can thus be detected when the ambient temperature is increasing or decreasing (Bulanon et al., 2008; Stajnko et al., 2004). Multispectral and hyperspectral cameras have also been used for fruit detection, allowing the acquisition of data at different bands of the electromagnetic spectrum (Okamoto and Lee, 2009; Sa et al., 2016; Safren et al., 2007; Zhang et al., 2015). However, like RGB cameras, thermal, multispectral and hyperspectral cameras only provide 2D information.

More recently, LiDAR (Light Detection and Ranging) systems have been introduced in agriculture to obtain 3D models of crops (Escolà et al., 2017; Rosell Polo et al., 2009; Sanz et al., 2018). This sensor works according to the time-of-flight principle (ToF), measuring distances to the objects by computing the time required by a laser pulse to complete the round trip between sensor and target. Besides the geometrical information (3D point clouds), this sensor also provides the amount of light backscattered by the scene (related with the reflectance). In this respect, the authors have shown in a recent study

(Gené-Mola et al., 2018) that some fruits, like Fuji apples, have higher reflectance than leaves and trunks, reporting an 85% detection success rate when using the reflectance capabilities of a LiDAR sensor.

Another technology derived from previous ones, and also used in crop monitoring are the RGB-D or depth cameras (Rosell-Polo et al., 2017, 2015). These sensors provide 3D information with color data, allowing the detection and subsequent 3D localization of the fruit. The operating principle can be based on stereo triangulation (Font et al., 2014; Wang et al., 2017) or on a combination of an RGB and a depth sensor, either based on structured light (Nguyen et al., 2016) or on ToF (Barnea et al., 2016; Gongal et al., 2018). Similarly to LiDAR sensors, RGB-D systems based on the ToF principle provide the amount of light backscattered by the scene, which can be related to the reflectance after range correction and sensor calibration (Rodríguez-Gonzálvez et al., 2016). This radiometric capability has been applied to face detection (Chhokra et al., 2018) by using the backscattered IR intensity image (without range correction) as an additional channel (RGB-DI). However, to the best of the authors' knowledge, no previous object detection work has used range-corrected intensity data (proportional to reflectance). The use of this additional information would be of interest in fruit detection, since the reflectance of some fruit varieties is higher than background reflectance (Gené-Mola et al., 2018).

Regarding the processing techniques used for fruit detection, most previous works have used traditional hand-crafted features to encode the data acquired with different sensors and infer fruit location. More recently, the introduction of deep neural networks has led to remarkable progress in object recognition and, therefore, in fruit detection. The object detection network Faster R-CNN (Ren et al., 2017) is the most commonly used for fruit detection (Bargoti and Underwood, 2017a; Gan et al., 2018; Sa et al., 2016; Stein et al., 2016). Although other state-of-the-art networks are computationally more efficient (Liu et al., 2016; Redmon and Farhadi, 2017), real-time inference is not normally a requirement in fruit detection, so Faster R-CNN is often chosen due to its better performance with small objects. The main drawback of using convolutional neural networks is that they require a large amount of labelled data. As pointed out in previous studies (Hameed et al., 2018), the lack of substantial datasets is a barrier for exploring emerging sensors that could be useful for fruit detection.

3. Materials and Methods

3.1. Theoretical basis

As previously introduced, the operation of the Kinect v2 sensor is based on the ToF principle. Thus, the received power coming from an object located at a distance R is given by the elastic LiDAR equation (Höfle and Pfeifer, 2007; Rodríguez-Gonzálvez et al., 2016):

$$P_r = \frac{P_t A \rho}{\pi R^2} \eta_{sys} \eta_{atm} \cos\theta \,, \tag{1}$$

where P_t is the emitted power, A is the receiving area, ρ is the object reflectance, η_{sys} is the optical efficiency of the instrument, η_{atm} accounts for atmospheric absorption and scattering, and θ is the incidence angle. The η_{atm} is assumed to be equal to unity due to the short working range of the Kinect sensor. The received power P_r is range corrected in order to compare returns coming from different distances:

$$P_r R^2 = K \rho \cos\theta \ , \tag{2}$$

where K is the system constant that groups the instrument parameters. Rodríguez-Gonzálvez et al. (2016) showed that there is a linear relationship between the digitized intensity E provided by the Kinect v2 and the received power:

$$E = a \cdot P_r + b \tag{3}$$

where a is the gain and b is the offset. From Eq. (2) and Eq. (3) it is found that the range-corrected signal S depends on the reflectance ρ as follows,

$$S = E R^2 \propto \rho \cos\theta \tag{4}$$

where $R^2 = x^2 + y^2 + z^2$, with $[x, y, z]$ the Cartesian coordinates of each point in the 3D cloud with respect to the sensor.

3.2. KFuji RGB-DS dataset

3.2.1. Data acquisition

Data were acquired in a commercial Fuji apple orchard (*Malus domestica* Borkh. cv. Fuji) located in Agramunt, Catalonia, Spain. The images were taken on September 25-28th of 2017, three weeks before harvesting, at BBCH (Biologische Bundesanstalt, Bundessortenamt und CHemische Industrie) phenological growth stage 85 (Meier, 2001).

The data acquisition equipment consisted of two RGB-D cameras mounted on a mobile platform at heights of 1 m and 3 m, respectively (Fig. 1) in order to capture data from all the tree height. The RGB-D sensors used were two Microsoft Kinect v2, which incorporate an RGB camera and a depth sensor that works according to the ToF principle. This sensor provides 3 different types of data: a color image, a depth image that can be used to generate a 3D point cloud of the scene, and the received IR backscattered intensity. Specific software written in C# was developed to collect and save data automatically. The software generates a 3D point cloud for each capture, with RGB and backscattered intensity data for each point, and saves it jointly with the raw RGB image. All captures were carried out during the night, using artificial

lighting, since performance of the depth sensor drops under direct sunlight exposure (Rosell-Polo et al., 2015). Table 1 summarizes the specifications of the sensor and the platform used for data acquisition.

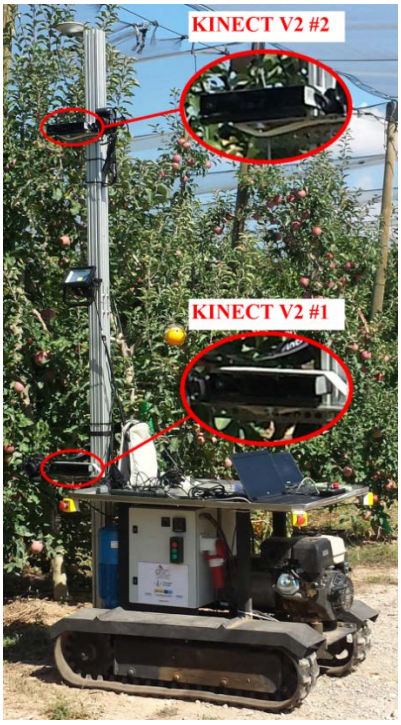

Fig. 1. View of the acquisition equipment showing the Kinect v2 sensors mounted on the mobile platform.

Table 1. Measurement equipment specifications.

RGB-D sensor	Manufacturer and model	Microsoft Kinect v2
	RGB channel resolution (pixels)	1920 x 1080
	RGB channel field-of-view (FOV)	84.1º x 53.8º
	IR and Depth channel resolution (pixels)	512 x 424
	IR and Depth channel FOV	70º x 60º
	Working range (m)	0.5 - 8
Mobile Platform	Developer	GRAP-UdL-AT research group
	Forward speed (km/h)	0.5 (manually adjustable)
	Sensors height (m)	1 - 3

3.2.2. Data preparation

Once the data were collected, for each capture it was obtained a 3D point cloud (with RGB and backscattered intensity information) and the corresponding raw RGB image. Captures were processed separately. Depending on the application where this methodology is used, apples appearing in the overlapped parts of images should be addressed. For instance, for

yield estimation, images should be registered in order to count the apples appearing in the overlapped parts only one time. However, this is not the goal of this work.

A pre-processing was carried out to prepare these data as input data of the convolutional neural network. Data preparation included range-correction of the backscattered signal, 2D projection of the 3D point cloud, and image registration between range-corrected intensity and raw RGB images. Fig. 2 illustrates a flowchart of the data preparation steps.

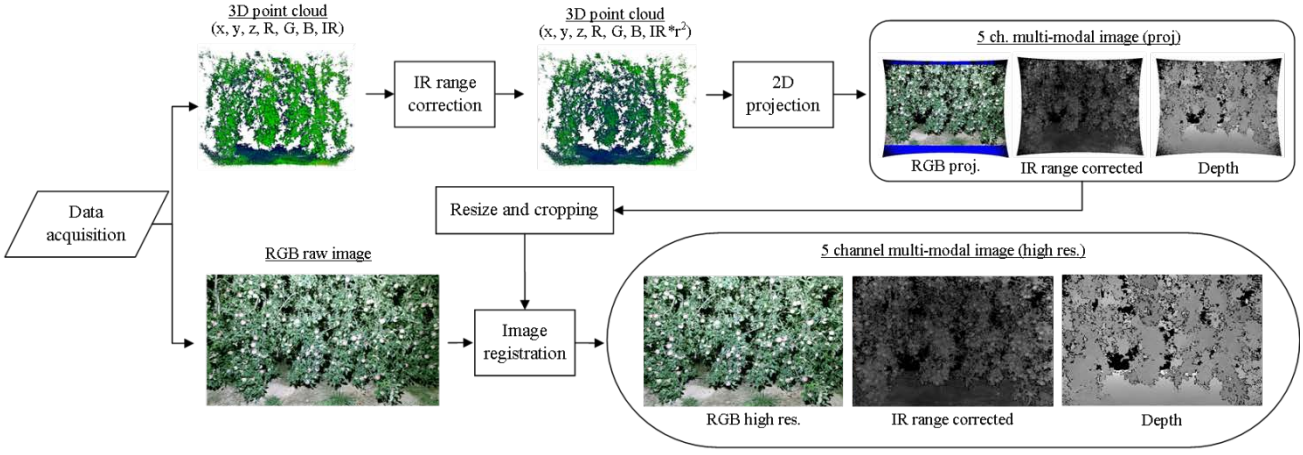

Fig. 2. Data preparation diagram. For each frame, the sensor provides a 3D point cloud with RGB and backscattered intensity data, and a raw RGB image. Firstly, the intensity signal is range-corrected. Then, the 3D point cloud is projected onto a 2D plane parallel to the sensor, generating the range-corrected and depth images. Finally, the projected images are resized and cropped in order to register them with the RGB raw image. 5-channel multi-modal images of the scene were obtained, with RGB channels in high resolution.

The range correction of the backscattered signal was performed as described in Section 3.1. After obtaining the 3D point cloud with range-corrected intensity data, a perspective projection onto a plane parallel to the sensor was carried out, generating the corresponding RGB projected, range-corrected intensity and depth images. Since the vertical field-of-view (FOV) of the depth sensor is larger than the vertical FOV of the RGB camera, the top and bottom parts, where no RGB information is given, were cut (blue regions in the RGB proj. image of Fig. 2). This step was the responsible of having a different image aspect ratio that the original 512/424. In order to work with an RGB image with higher resolution than that of the IR image, a registration between the RGB raw image and the projected images is required. To do so, projected images were resized (bicubic interpolation) to 1600 x 1080 pixels (px) to achieve the same vertical size as the RGB high resolution image. Finally, an image registration was performed in order to have correspondence between all images, obtaining a 5-channel multi-modal image (with RGB in high resolution) where each pixel has information from 3 modalities: color (RGB), range-corrected intensity and depth (Fig. 2). Hereinafter, the RGB image obtained after the point cloud projection is denoted as RGB_p, while the RGB image obtained after registering the raw RGB image is denoted as RGB_{hr}. In order to have similar mean and variance between channels, the range-corrected intensity and depth images were

normalized between 0 and 255 -like RGB images-. This normalization is desirable to ensure fast convergence of the network. The RGB channels were saved in 8-bit images while the range-corrected intensity and depth images were stored in 64-bits to avoid data precision loss.

Ground truth fruit locations were manually annotated using the Pychet Labeller toolbox (Bargoti, 2016), labeling a total of 12,839 apples in all the dataset. Due to the large number of fruits per image (more than 100 fruits/image), and taking into account that fruit size (44 ± 6 px in diameter) is relatively small with respect to image size (1600 x 1080 px), each capture was divided into 9 sub-images of 548 x 373 px, with an overlap of 20 px between sub-images to avoid the partially split of fruits at the boundaries in different partitions (Fig. 3).

Fig. 3. Image sub-division. Each raw image was divided into 9 sub-images to achieve a better relation between apple and image size.

In total, the data set is composed of 967 sub-images, split into training, validation, and test sets as shown in Table 2. Some examples of the multi-modal sub-images used in the training dataset are shown in Fig. 4. Due to further quantization for representation, fruits cannot be seen in depth images. The KFuji RGB-DS dataset with corresponding annotations has been made publicly available at *www.grap.udl.cat/en/publications/datasets.html.*

Table 2. Dataset configuration.

Raw image	Sub-image	Fruit size	Training	Validation	Test	No. of fruits (all dataset)
1600x1080 px	548x373 px	44 ± 6 px	619 (64%)	155 (16%)	193 (20%)	12.839

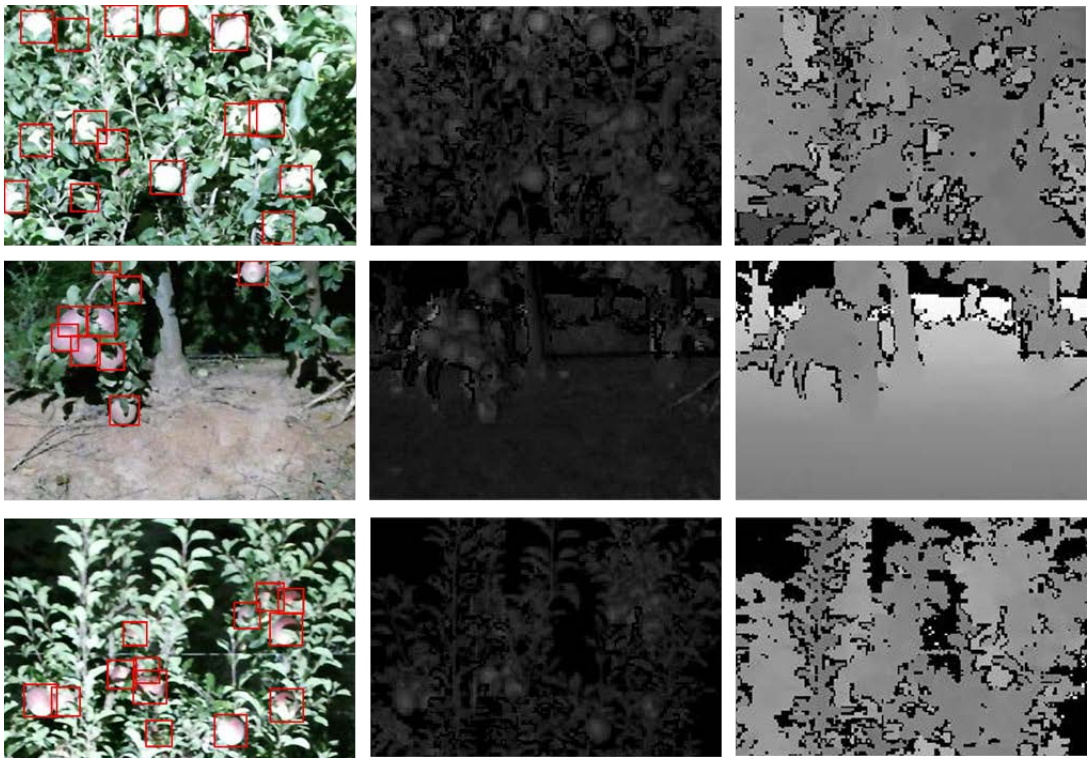

Fig. 4. Sample of 3 multi-modal images extracted from training dataset and their associated fruit location ground truth (red bounding boxes). First column corresponds to RGB_{hr}, second column to S and the third column to D channel.

3.3. Experiments

The Faster R-CNN object detection network (Ren et al., 2017) was used in this work as fruit detector. The choice of Faster R-CNN allows a comparison of the performance of our methodology with previous works that also used Faster R-CNN for fruit detection (Bargoti and Underwood, 2017a; Gan et al., 2018; Sa et al., 2016; Stein et al., 2016).

Faster R-CNN was originally developed to detect objects in color images. The original work (Ren et al., 2017) tested the network with PASCAL VOC (Everingham et al., 2010) and COCO (Lin et al., 2014) datasets, reporting a mean average precision (mAP) of 78.8% and 42.7% for the VOC 2007 and COCO test sets, respectively. Faster R-CNN is composed of two modules: (1) a region proposal network (RPN), to identify promising regions of interest (ROIs) that are likely to contain an object; (2) a classification network, which classifies the regions proposed. Both parts share the first convolutional layers, making it a fast object detector. The RPN uses the feature maps produced by the first convolutional layers to produce promising ROIs by means of a series of convolutional and fully connected layers. The RPN output is then used to crop out corresponding regions from the feature maps produced by the first convolutional layers (crop pooling). The regions produced by crop pooling are then passed through a classification network and a regressor to predict the probability of a ROI being apple or background and refine the ROI. Fig. 5 illustrates a diagram of the implemented Faster R-CNN network.

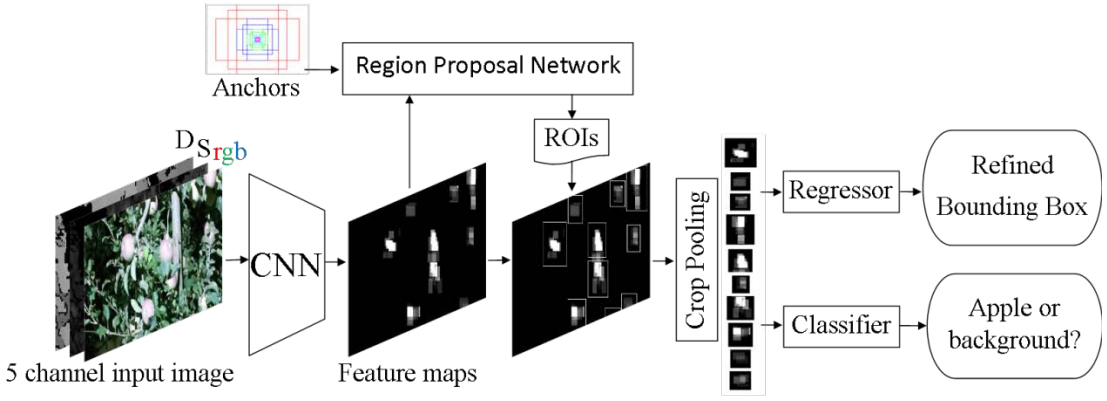

Fig. 5. Diagram of the implemented Faster R-CNN. The main modifications to the original Faster R-CNN are the multi-modal input and the anchor scales.

In this work, the first convolutional layers uses the VGG-16 model (Simonyan and Zisserman, 2014) pre-trained with ImageNet dataset (Deng et al., 2009) and fine-tuned with our training dataset. The original implementation of Faster R-CNN was modified to make it suitable for our dataset. The main modifications to the original Faster R-CNN done in this work are: (1) multi-modal input and (2) region proposal adaptation.

Regarding the multi-modal input, since the original implementation of Faster R-CNN uses color images, the input layer was modified to work with 5-channel images. Due to these additional channels, filters from the first convolutional layer increase in depth (from 3 to 5), which implies that more weights must be initialized. Thus, after loading pre-trained weights in the network, additional weights corresponding to channels D and S were randomly initialized.

To generate region proposals, the RPN evaluates different boxes in each position of the image with a stride of 16 pixels. The different types of boxes evaluated are called anchors and are characterized by their scale (box area) and the aspect ratio. The original implementation of Faster R-CNN proposed 3 anchor scales of 8, 16 and 32 -corresponding to box areas of 128^2, 256^2 and 512^2 pixels-, and 3 aspect ratios of 1:1, 1:2 and 2:1. Since the presented dataset has smaller objects than datasets tested in Ren et al. (2017), a study of the optimal anchor scales and aspect ratios was carried out. Besides the anchors proposed in the original paper (8, 16, 32), smaller anchor scales were also tested (2 and 4). The aspect ratios used in this study were the same as those used in the original implementation (1:1, 1:2 and 2:1), however, two different configurations were tested: only considering aspect ratio of 1:1, and combining the three aspect ratios. Fig. 6 illustrates the anchors tested in this work compared with the image size. Although using input images of 548 x 373 px, the network resizes the input images to 600 px on the shortest side. For this reason, the shortest side has 600 px instead of 373 px.

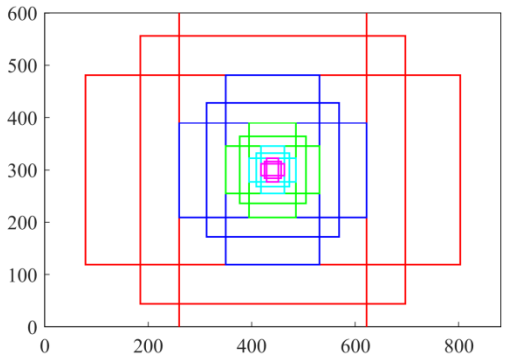

Fig. 6. Anchors tested compared with the image size. Five anchor scales were tested: 2 (magenta), 4 (cyan), 8 (green), 16 (blue) and 32 (red). For each anchor, three different aspect ratios were used: 0.5, 1 and 2.

To evaluate the performance, average precision (AP), precision, recall and F1-score metrics are reported for the test dataset. Predictions were considered as true positive if the intersection over union (IoU) between prediction and ground truth bounding boxes was greater than 0.5. The network was implemented in PyTorch framework (Paszke et al., 2017) and has been made publicly available at *www.grap.udl.cat/en/publications/datasets.html.*

4. Results and discussion

This section presents a qualitative and quantitative evaluation of the proposed fruit detection methodology, assessing the performance when using different image modalities provided by Kinect v2, and studying the optimal anchor scale configurations proposed in the RPN.

To study the usefulness of different image modalities (RGB, S and D), different Faster R-CNN models were trained using each modality separately as well as combinations thereof. This study was carried out using projected color images RGB_p since they have the same resolution as S and D images, to enable a comparison between modalities performed under the same conditions. Nevertheless, results with 5-channel multi-model images using RGB_{hr} are also provided to assess the potential of the sensor for fruit detection.

4.1. Training assessment

The network was trained end-to-end using the loss function proposed in Ren et al. (2017), which is comprised of the sum of a classification loss and a bounding box regression loss. Following Ren et al. (2017), the training loss function considered positive detections if IoU>0.7 and negative if IoU<0.3, while anchors that are neither positive nor negative do not contribute to the loss function. Adam optimizer with a learning rate of 0.0001 was used to update network weights iteratively, performing a total of 309 training (38 validation) iterations per epoch with a batch size of 4 images. Data augmentation was performed with left-right flipping to expand the variability of the training dataset. A validation set was

used to evaluate the training after each epoch to check model generalization and identify if it starts to overfit. The number of images used for training, validation and test were 619, 155 and 193, respectively. Fig. 7 shows the loss function for training and validation sets using different image modalities. By comparing models where RGB_p was used, it can be seen how when only using color modality (plotted in cyan) the model starts to overfit earlier than with the addition of S and D channels (plotted in orange and green). Therefore, the use of S and D channels allowed model training during more iterations without overfitting. With respect to training curves, the loss function archived lower values when using only color images. However, the opposite occurred with validation losses, with the best validation loss achieved by combining all modalities. This is a consequence of early overfitting of the RGB_p model, and, from that, it was concluded that the S and D channels helped model generalization, with better results obtained on the validation dataset when using 5-channel multi-modal images. On the other hand, when comparing the performance of using RGB_p or RGB_{hr} images, training and validation loss functions showed an important improvement when RGB_{hr} images were used. This improvement increased when adding S and D channels to RGB_{hr}, although not in the same proportion as multi-modal images with RGB_p. This suggests that if future RGB-D sensors had depth sensors with higher resolution (similar to color cameras), detection performance could be improved even further.

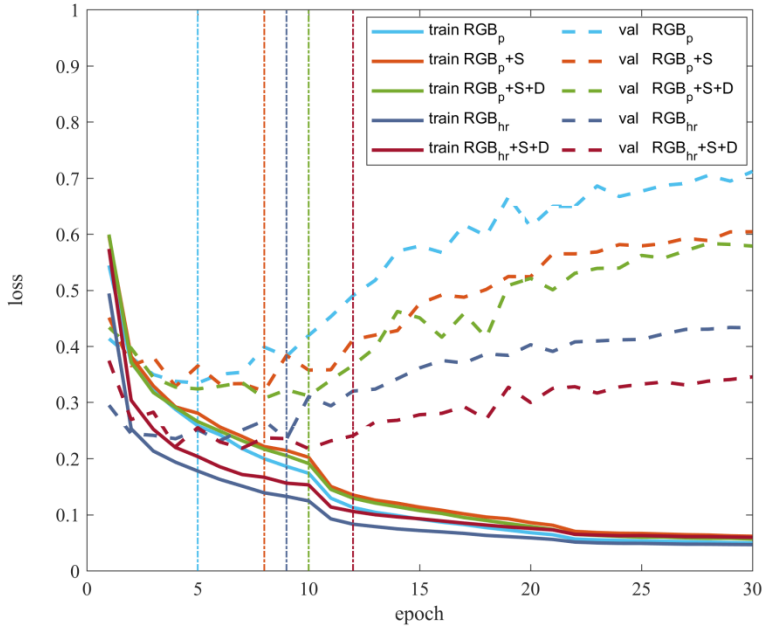

Fig. 7. Training and validation (val) losses depending on the number of training epochs. Loss function was computed following Ren et al. (2017). In cyan, the evolution of training and validation losses is plotted using RGB projected images (RGB_p). Orange data refer to projected images with range-corrected signal intensity (S) data, and green data when adding depth (D) information as well. Shown in blue are the results corresponding to use of high resolution RGB (RGB_{hr} -registered row image-). Finally, results of multi-modal data with RGB_{hr}, range-corrected signal and depth images are plotted in red. Vertical lines mark the epoch where each model starts to overfit, which is the epoch in which test results are reported.

4.2. Anchor optimization

This section evaluates the performance of Faster R-CNN with multi-modal images (RGB_{hr}+S+D) depending on the anchor scales used in the RPN. The original paper of Faster R-CNN (Ren et al., 2017) used anchor scales of 8-16-32, but it mentions that the anchor scales used were not specifically chosen for a particular dataset. The present work was evaluated on a very different dataset (with small spherical objects) from the one used in the original Faster R-CNN work. Therefore, the behavior of the network using anchor scales of 2, 4, 8, 16 and 32 and aspect ratios of 1:2, 1:1, 2:2 was analyzed.

Table 3 presents the results obtained when using different anchor scales and aspect ratios in terms of AP. Comparing anchor scales configurations, the worst performances were achieved using anchor scales of 16 and 32, while other configurations performed similarly, with an AP ranging between 93.4% (using anchor scale of 8 and aspect ratios of 1:2, 1:1 and 2:1), and 94.8% (using anchor scales of 4 and aspect ratios of 1:1). Regarding the anchor aspect ratios, best results were obtained only using squared anchors (anchor ratios of 1:1). This responds the fact that fruits are spherical. As for computational efficiency, Table 3 shows that frame rate slightly decreases when combining different anchor scales or aspect ratios. This fact is due to the number of convolutional operations in the RPN increases. However, since the number of object proposals was limited to 100 in all cases (as suggested in Sa et al. 2016) the computational efficiency do not show important differences. From these results, the following sections use anchor scales of 4 and aspect ratios of 1:1, being the configuration that showed the best performance.

Table 3. Fruit detection results on RGB_{hr}+S+D test set using different anchor scales and ratios.

	Anchor aspect ratios			
	[1:2 , 1:1, 2:1]		[1:1]	
Anchor scales	AP (%)	frames/s	AP (%)	frames/s
8-16-32 (Ren et al., 2017)	93.1	12.7	92.8	12.9
2	93.6	12.7	94.0	13.1
4	93.5	12.6	**94.8**	13.6
8	93.4	12.4	93.7	13.3
16	91.3	13.0	86.6	13.2
32	79.9	12.9	86.4	13.0
2-4-8	94.1	12.7	94.4	12.4
4-8-16	94.1	12.6	93.5	12.1
2-4-8-16	93.1	12.8	94.1	12.8

Fig. 8 illustrates some fruit detection examples using anchor scales of 4 and aspect ratios of 1:1. Images were selected to show cases where the network succeeds or fails, so that the illustrated examples correspond to the four best (first column), four intermediate (second column) and four worst (third column) scored images from the test dataset. As can be seen, most of the false positives correspond to image regions that are very similar to apples or to real apples that were not labelled because of human errors when labelling. On the other hand, most of the false negatives correspond to highly occluded apples and to apples that were cut by the borders of the image.

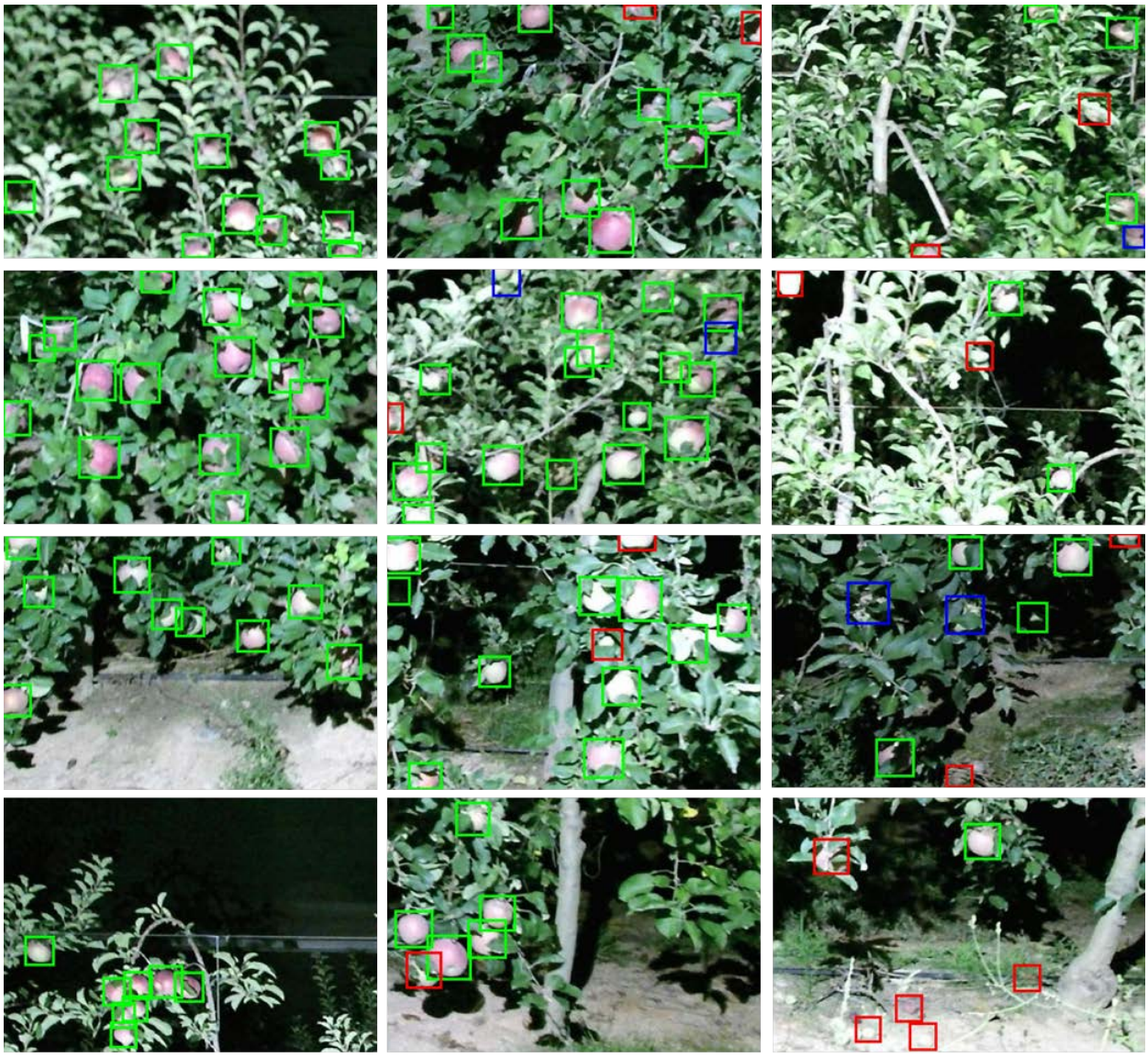

Fig. 8. Fruit detection results on RGB_{hr}+S+D test set using anchor scales of 4, 8 and 16. True positive detections are shown in green squares, false positives in red and false negatives in blue. Images are ordered according to their F1-score. Column (a) contains examples of the best detection results, F1-score = 1. Column (b) contains the four intermediate scored images of the test set, corresponding to an F1-score =0.91. Column (c) shows the worst detections, ordered from bottom to top with an F1-score = [0.23, 0.67, 0.67, 0.67].

4.3. Test results from different modalities

Regarding the test set, Table 4 presents fruit detection results obtained from different input image modalities. The performance of Faster R-CNN using each input type was evaluated in terms of Precision (P), Recall (R), F1-score, AP and number of inferred images per second (processing on a GeForce GTX TITAN X GPU). A confidence threshold of 0.85 was selected from Precision and Recall curves (Fig. 9). The number of training epoch is also given. This number was chosen from training and validation loss curves, selecting the last epoch that did not present overfit (vertical lines in Fig. 7). Fig. 10 shows graphically the fruit detection of three selected images from the test set using different input modalities (RGB_p, S, D, RGB_p+S and RGB_p+S+D). True positives are shown in green, false positives in red and false negatives in blue.

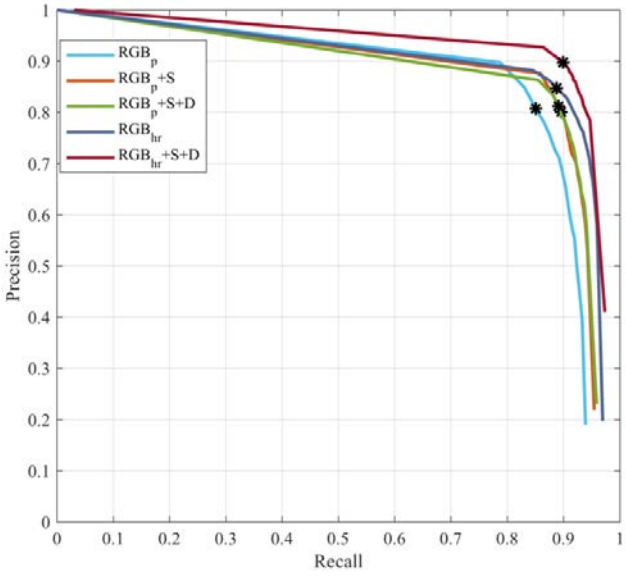

Fig. 9. Precision and Recall curves obtained for different image modalities. Black asterisks correspond to the working points with the selected confidence threshold of 0.85.

Table 4. Fruit detection results from test dataset using different image modalities.

Channels	Epoch	P	R	F1-score	AP (%)	frames/s
RGB_p	5	0.808	0.851	0.829	88.7	13.4
S	4	0.848	0.768	0.806	85.9	13.5
D	7	0.699	0.582	0.635	61.3	13.4
RGB_p+S	8	0.887	0.827	0.856	89.8	12.9
RGB_p+D	9	0.802	0.848	0.824	88.0	13.7
S+D	9	0.731	0.821	0.774	84.6	13.4
RGB_p+S+D	10	**0.869**	**0.864**	**0.866**	**91.2**	13.1
RGB_{hr}	9	0.847	0.888	0.867	92.7	12.9
RGB_{hr}+S+D	12	**0.897**	**0.899**	**0.898**	**94.8**	13.6

Comparing results from single modality images (Table 4, rows 1-3), color images gave the best performance with an F1-score of 0.829 and an AP of 88.7%, followed by the range-corrected intensity image with an F1-score of 0.806 and an AP of 85.9%. Note that, although range-corrected intensity images have never been used for fruit detection, the results using this modality are comparable with other state-of-the-art methods. The least valuable modality was the Depth channel which was only able to detect highly exposed (non-occluded) apples, as can be seen in Fig. 10. Better results were obtained when combining different modalities (multi-modal images), achieving an F1-score of 0.866 and an AP of 91.2% when all channels were used. The most important benefit of adding S and D was found in the Precision metric, which rose from 0.808 (RGB_p) to 0.869 (RGB_p+S+D), although Recall and AP also improved albeit in smaller percentages. This means that range-corrected intensity and depth images help to reduce false positives. Real examples of this effect can be found in Fig. 10, where, when comparing results before and after using S and D channels, a reduction in false positives is observed. Another advantage of using the S channel was found when detecting fruits in shadowed regions, where the RGB image presents a dark non-colored region whereas the S channel shows high intensities. This occurs in Fig. 10b, where an apple in a shadowed region was not detected using RGB_p, but was detected using the S channel.

Finally, as was expected, the best performance was achieved using multi-modal images with RGB_{hr}, S and D, reporting an F1-score of 0. 898 and a AP of 94.8%. Regarding the computational efficiency of the neural network, the number of inferred images per second did not present any relation with the number of channels used. This is because the addition of channels only affects the number of operations on the first layer, which is insignificant with respect to the whole network.

Although it is difficult to compare methodologies tested with different datasets, results shows similar performance to other fruit detection works based on neural networks, such as Bargoti and Underwood (2017), Gan et al. (2018) and Sa et al. (2016) which reported F1-scores between 0.838 and 0.929 (using less restrictive IoU thresholds than the present work). However, the use of RGB-D sensors has the advantage that, although detecting fruits in 2D images, it is straightforward to infer the 3D location of each detection.

The main limitation of the proposed methodology is that the working conditions are restricted to low illuminance levels. However, we expect that future sensors could solve this limitation. For instance, LiDAR-based sensors are already able to build 3D models with reflectance data in natural lighting conditions. In addition, convolutional neural networks have shown good performances with wide enough datasets that contain different illumination conditions (Amara et al., 2017; Chen et al., 2017; Rahnemoonfar and Sheppard, 2017). From that, we expect that if future RGB-D sensors would not be influenced by high illumination levels, the methodology proposed could be used in daylight.

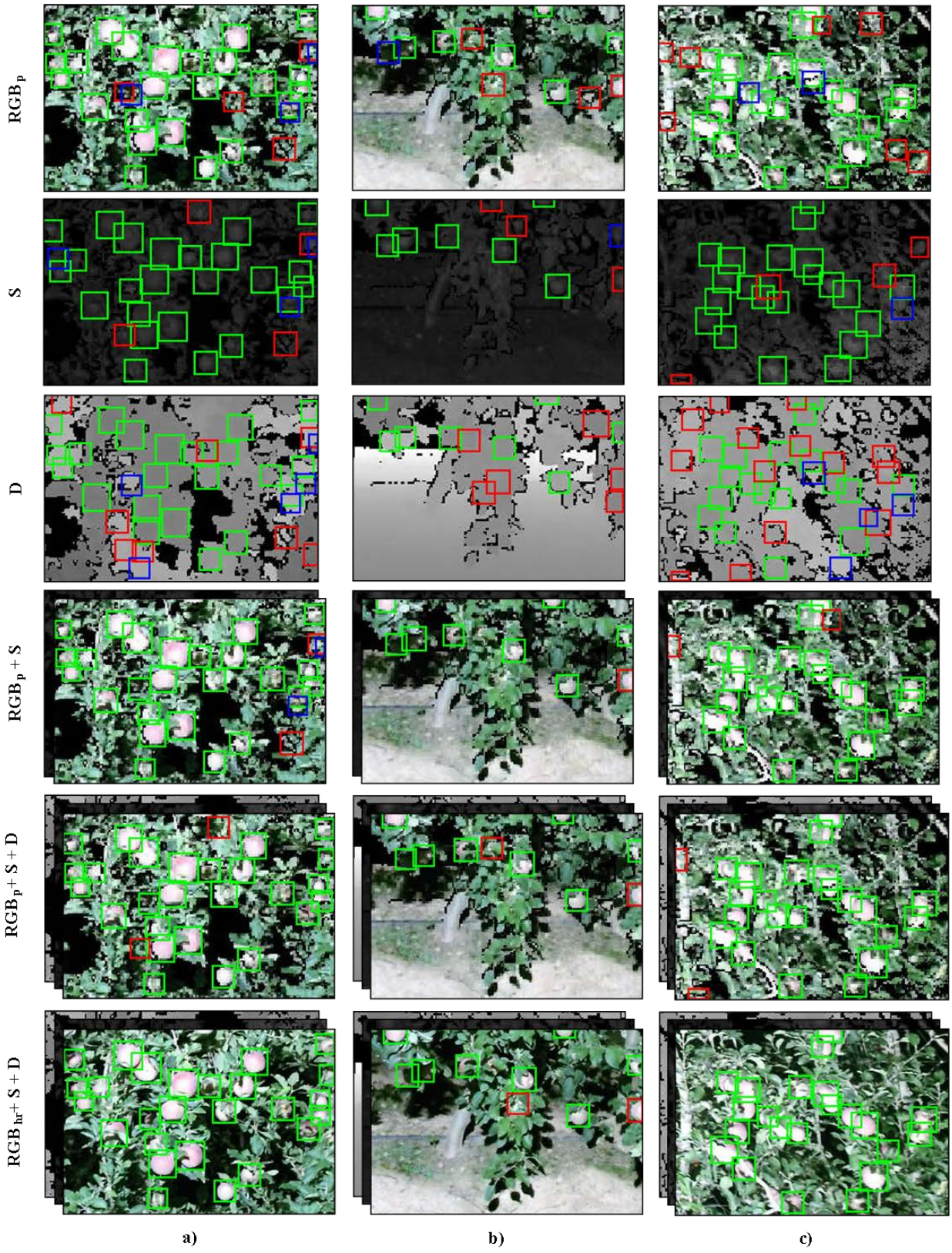

Fig. 10. Selected examples of fruit detection results to show the effect of adding range-corrected signal intensity (S) and depth (D) information. For each sample a), b) and c), six different fruit detection results are shown depending on the input data type: RGB_p (first row), S (second row), D (third row), multi-modal RGB_p and S (fourth row), using all modalities RGB_p, S and D (fifth row), and using all modalities with high resolution image RGB_{hr}, S and D (last row). True positive detections are shown in green, false positives in red and false negatives in blue.

5. Conclusions

This work presents a novel methodology for fruit detection using RGB-D sensors, taking advantage of their radiometric capabilities. Multi-modal images were built using data provided by Microsoft's Kinect v2. To do so, a range correction of the backscattered intensity signal was carried out to overcome signal attenuation (R^{-2} dependence). Then, a registration between different channels was performed, obtaining images with 3 modalities: color (RGB), depth (D), and range-corrected intensity (S). The KFuji RGB-DS dataset and the corresponding annotations have been made publicly available, being the first dataset for fruit detection that contains RGB, depth and range-corrected intensity channels. The Faster R-CNN object detection network was used to evaluate the usefulness of fusing all modalities. Results show an improvement of 4.46% in F1-score when all modalities were used -from 0.829 (RGB_p) to 0.866 (RGB_p-D-S)-. This entails an advance in the field of fruit detection, since the results are comparable to other fruit detection methodologies retrieved from the state of the art, with the additional advantage that, by using RGB-D sensors, it is possible to infer the 3D position of each detection. The optimum anchor scales used in the region proposal network were also analyzed. It is concluded that, for KFuji RGB-DS dataset where fruits are spherical and small with respect to the image size, the optimum configuration were anchor scales of 4 and aspect ratios of 1:1. The main limitation of using RGB-D is that the performance of the depth sensor drops under direct sunlight. Future works will include 3D fruit localization by projecting the 2D fruit detection onto the 3D world using the depth channel data as well as collecting data of different fruit varieties and at different growth stages.

Acknowledgements

This work was partly funded by the Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement de la Generalitat de Catalunya, the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund (ERDF) under Grants 2017 SGR 646, AGL2013-48297-C2-2-R and MALEGRA, TEC2016-75976-R. The Spanish Ministry of Education is thanked for Mr. J. Gené's pre-doctoral fellowships (FPU15/03355). We would also like to thank Nufri and Vicens Maquinària Agrícola S.A. for their support during data acquisition, and Adria Carbó for his assistance in Faster R-CNN implementation.

References

Amara, J., Bouaziz, B., Algergawy, A., 2017. A Deep Learning-based Approach for Banana Leaf Diseases Classification. Btw 79–88.

Auat Cheein, F.A., Carelli, R., 2013. Agricultural robotics: Unmanned robotic service units in agricultural tasks. IEEE Ind. Electron. Mag. 7, 48–58. doi:10.1109/MIE.2013.2252957

Bargoti, S., 2016. Pychet Labeller. Available online: https://github.com/acfr/pychetlabeller.

Bargoti, S., Underwood, J., 2017a. Deep Fruit Detection in Orchards. 2017 IEEE Int. Conf. Robot. Autom. 3626–3633.

Bargoti, S., Underwood, J.P., 2017b. Image Segmentation for Fruit Detection and Yield Estimation in Apple Orchards. J. F. Robot. 00, 1–22. doi:10.1002/rob.21699

Barnea, E., Mairon, R., Ben-Shahar, O., 2016. Colour-agnostic shape-based 3D fruit detection for crop harvesting robots. Biosyst. Eng. 146, 57–70. doi:10.1016/j.biosystemseng.2016.01.013

Bulanon, D.M., Burks, T.F., Alchanatis, V., 2008. Study on temporal variation in citrus canopy using thermal imaging for citrus fruit detection. Biosyst. Eng. 101, 161–171. doi:10.1016/j.biosystemseng.2008.08.002

Chen, S.W., Shivakumar, S.S., Dcunha, S., Das, J., Okon, E., Qu, C., Taylor, C.J., Kumar, V., 2017. Counting Apples and Oranges With Deep Learning: A Data-Driven Approach. IEEE Robot. Autom. Lett. 2, 781–788. doi:10.1109/LRA.2017.2651944

Chhokra, P., Chowdhury, A., Goswami, G., Vatsa, M., Singh, R., 2018. Unconstrained Kinect video face database. Inf. Fusion 44, 113–125. doi:10.1016/j.inffus.2017.09.002

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition. doi:10.1109/CVPRW.2009.5206848

Escolà, A., Martinez-Casasnovas, J.A., Rufat, J., Arno, J., Arbones, A., Sebe, F., Pascual, M., Gregorio, E., Rosell-Polo, J.R., 2017. Mobile terrestrial laser scanner applications in precision fruticulture/horticulture and tools to extract information from canopy point clouds. Precis. Agric. 18, 111–132. doi:10.1007/s11119-016-9474-5

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2010. The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. doi:10.1007/s11263-009-0275-4

Font, D., Pallejà, T., Tresanchez, M., Runcan, D., Moreno, J., Martínez, D., Teixidó, M., Palacín, J., 2014. A proposal for automatic fruit harvesting by combining a low cost stereovision camera and a robotic arm. Sensors (Switzerland) 14, 11557–11579. doi:10.3390/s140711557

Gan, H., Lee, W.S., Alchanatis, V., Ehsani, R., Schueller, J.K., 2018. Immature green citrus fruit detection using color and thermal images. Comput. Electron. Agric. 152, 117–125. doi:10.1016/j.compag.2018.07.011

Gené-Mola, J., Gregorio, E., Guevara, J., Auat, F., Escolà, A., Morros, J.-R., Rosell-Polo, J.R., 2018. Fruit Detection Using Mobile Terrestrial Laser Scanning, in: EurAgEng 2018 Conference.

Gongal, A., Amatya, S., Karkee, M., Zhang, Q., Lewis, K., 2015. Sensors and systems for fruit detection and localization: A review. Comput. Electron. Agric. 116, 8–19. doi:10.1016/j.compag.2015.05.021

Gongal, A., Karkee, M., Amatya, S., 2018. Apple fruit size estimation using a 3D machine vision system. Inf. Process. Agric. 5, 498–503. doi:10.1016/j.inpa.2018.06.002

Hameed, K., Chai, D., Rassau, A., 2018. A comprehensive review of fruit and vegetable classification techniques. Image Vis. Comput. #pagerange#. doi:10.1016/J.IMAVIS.2018.09.016

Höfle, B., Pfeifer, N., 2007. Correction of laser scanning intensity data: Data and model-driven approaches. ISPRS J. Photogramm. Remote Sens. 62, 415–433. doi:10.1016/j.isprsjprs.2007.05.008

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common objects in context, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). doi:10.1007/978-3-319-10602-1_48

Linker, R., 2017. A procedure for estimating the number of green mature apples in night-time orchard images using light distribution and its application to yield estimation. Precis. Agric. 18, 59–75. doi:10.1007/s11119-016-9467-4

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C., 2016. SSD: Single shot multibox detector, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). doi:10.1007/978-3-319-46448-0_2

Maldonado, W., Barbosa, J.C., 2016. Automatic green fruit counting in orange trees using digital images. Comput. Electron. Agric. 127, 572–581. doi:10.1016/j.compag.2016.07.023

Meier, U., 2001. Growth stages of mono- and dicotyledonous plants, BBCH Monograph. doi:10.5073/bbch0515

Nguyen, T.T., Vandevoorde, K., Wouters, N., Kayacan, E., De Baerdemaeker, J.G., Saeys, W., 2016. Detection of red and bicoloured apples on tree with an RGB-D camera. Biosyst. Eng. 146, 33–44. doi:10.1016/j.biosystemseng.2016.01.007

Nuske, S., Wilshusen, K., Achar, S., Yoder, L., Singh, S., 2014. Automated visual yield estimation in vineyards, in: Journal of Field Robotics. doi:10.1002/rob.21541

Okamoto, H., Lee, W.S., 2009. Green citrus detection using hyperspectral imaging. Comput. Electron. Agric. 66, 201–208. doi:10.1016/j.compag.2009.02.004

Paszke, A., Chanan, G., Lin, Z., Gross, S., Yang, E., Antiga, L., Devito, Z., 2017. Automatic differentiation in PyTorch. Adv. Neural Inf. Process. Syst. 30.

Rahnemoonfar, M., Sheppard, C., 2017. Deep count: Fruit counting based on deep simulated learning. Sensors (Switzerland) 17, 1–12. doi:10.3390/s17040905

Redmon, J., Farhadi, A., 2017. YOLO9000: Better, faster, stronger, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. doi:10.1109/CVPR.2017.690

Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans. Pattern Anal. Mach. Intell. 39, 1137–1149. doi:10.1109/TPAMI.2016.2577031

Rodríguez-Gonzálvez, P., Gonzalez-Aguilera, D., González-Jorge, H., Hernández-López, D., 2016. Low-Cost Reflectance-Based Method for the Radiometric Calibration of Kinect 2. IEEE Sens. J. 16, 1975–1985. doi:10.1109/JSEN.2015.2508802

Rosell-Polo, J.R., Cheein, F.A., Gregorio, E., Andújar, D., Puigdomènech, L., Masip, J., Escolà, A., 2015. Advances in Structured Light Sensors Applications in Precision Agriculture and Livestock Farming, in: Advances in Agronomy. doi:10.1016/bs.agron.2015.05.002

Rosell-Polo, J.R., Gregorio, E., Gene, J., Llorens, J., Torrent, X., Arno, J., Escola, A., 2017. Kinect v2 Sensor-based Mobile Terrestrial Laser Scanner for Agricultural Outdoor Applications. IEEE/ASME Trans. Mechatronics 1–1. doi:10.1109/TMECH.2017.2663436

Rosell Polo, J.R., Sanz, R., Llorens, J., Arnó, J., Escolà, A., Ribes-Dasi, M., Masip, J., Camp, F., Gràcia, F., Solanelles, F., Pallejà, T., Val, L., Planas, S., Gil, E., Palacín, J., 2009. A tractor-mounted scanning LIDAR for the non-destructive measurement of vegetative volume and surface area of tree-row plantations: A comparison with conventional destructive measurements. Biosyst. Eng. 102, 128–134. doi:10.1016/j.biosystemseng.2008.10.009

Sa, I., Ge, Z., Dayoub, F., Upcroft, B., Perez, T., McCool, C., 2016. DeepFruits: A Fruit Detection System Using Deep Neural Networks. Sensors 16, 1222. doi:10.3390/s16081222

Safren, O., Alchanatis, V., Ostrovsky, V., Levi, O., 2007. Detection of Green Apples in Hyperspectral Images of Apple-Tree Foliage Using machine Vision 50, 2303–2313.

Sanz, R., Llorens, J., Escolà, A., Arnó, J., Planas, S., Román, C., Rosell-Polo, J.R., 2018. LIDAR and non-LIDAR-based canopy parameters to estimate

the leaf area in fruit trees and vineyard. Agric. For. Meteorol. 260–261, 229–239. doi:10.1016/j.agrformet.2018.06.017

Siegel, K.R., Ali, M.K., Srinivasiah, A., Nugent, R.A., Narayan, K.M.V., 2014. Do we produce enough fruits and vegetables to meet global health need? PLoS One 9. doi:10.1371/journal.pone.0104059

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition 1–14. doi:10.1016/j.infsof.2008.09.005

Stajnko, D., Lakota, M., Hocevar, M., 2004. Estimation of number and diameter of apple fruits in an orchard during the growing season by thermal imaging. Comput. Electron. Agric. 42, 31–42. doi:10.1016/S0168-1699(03)00086-3

Stein, M., Bargoti, S., Underwood, J., 2016. Image Based Mango Fruit Detection, Localisation and Yield Estimation Using Multiple View Geometry. Sensors 16, 1915. doi:10.3390/s16111915

Underwood, J.P., Hung, C., Whelan, B., Sukkarieh, S., 2016. Mapping almond orchard canopy volume, flowers, fruit and yield using lidar and vision sensors. Comput. Electron. Agric. 130, 83–96. doi:10.1016/j.compag.2016.09.014

Uribeetxebarria, A., Daniele, E., Escolà, A., Arnó, J., Martínez-Casasnovas, J.A., 2018. Spatial variability in orchards after land transformation: Consequences for precision agriculture practices. Sci. Total Environ. doi:10.1016/j.scitotenv.2018.04.153

Wang, Z., Walsh, K.B., Verma, B., 2017. On-Tree Mango Fruit Size Estimation Using RGB-D Images. Sensors (Basel). 17. doi:10.3390/s17122738

Zhang, B., Huang, W., Wang, C., Gong, L., Zhao, C., Liu, C., Huang, D., 2015. Computer vision recognition of stem and calyx in apples using near-infrared linear-array structured light and 3D reconstruction. Biosyst. Eng. 139, 25–34. doi:10.1016/j.biosystemseng.2015.07.011

Zhao, C., Lee, W.S., He, D., 2016. Immature green citrus detection based on colour feature and sum of absolute transformed difference (SATD) using colour images in the citrus grove. Comput. Electron. Agric. 124, 243–253. doi:10.1016/j.compag.2016.04.009